

The implications of alternative splicing in the ENCODE protein complement

Michael L. Tress^{a,b}, Pier Luigi Martelli^c, Adam Frankish^d, Gabrielle A. Reeves^e, Jan Jaap Wesselink^a, Corin Yeats^f, Páll Ísólur Ólason^g, Mario Albrecht^h, Hedi Hegyiⁱ, Alejandro Giorgetti^j, Domenico Raimondo^j, Julien Lagarde^k, Roman A. Laskowski^e, Gonzalo López^a, Michael I. Sadowski^l, James D. Watson^e, Piero Fariselli^c, Ivan Rossi^c, Alinda Nagyⁱ, Wang Kai^g, Zenia Størling^g, Massimiliano Orsini^m, Yassen Assenov^h, Hagen Blankenburg^h, Carola Huthmacher^h, Fidel Ramírez^h, Andreas Schlicker^h, France Denoeud^k, Phil Jones^e, Samuel Kerrien^e, Sandra Orchard^e, Stylianos E. Antonarakisⁿ, Alexandre Reymond^o, Ewan Birney^e, Søren Brunak^g, Rita Casadio^c, Roderic Guigo^{k,p}, Jennifer Harrow^d, Henning Hermjakob^e, David T. Jones^l, Thomas Lengauer^h, Christine A. Orengo^f, László Patthyⁱ, Janet M. Thornton^{e,q}, Anna Tramontano^r, and Alfonso Valencia^{a,r}

^aStructural Computational Biology Programme, Spanish National Cancer Research Centre, E-28029 Madrid, Spain; ^bDepartment of Biology, University of Bologna, 33-40126 Bologna, Italy; ^cHAVANA Group, The Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA, United Kingdom; ^dEuropean Bioinformatics Institute, Hinxton, Cambridge CB10 1SD, United Kingdom; ^eDepartment of Biochemistry and Molecular Biology and ^fBioinformatics Unit, University College London, London WC1E 6BT, United Kingdom; ^gCenter for Biological Sequence Analysis, BioCentrum-DTU, DK-2800 Lyngby, Denmark; ^hMax Planck Institute for Informatics, 66123 Saarbrücken, Germany; ⁱBiological Research Center, Hungarian Academy of Sciences, 1113 Budapest, Hungary; ^jDepartment of Biochemical Sciences, University of Rome "La Sapienza," 2-00185 Rome, Italy; ^kResearch Unit on Biomedical Informatics, Institut Municipal d'Investigació Mèdica, E-8003 Barcelona, Spain; ^lCenter for Advanced Studies, Research and Development in Sardinia (CRS4), 09010 Pula, Italy; ^mDepartment of Genetic Medicine and Development, University of Geneva Medical School, 1211 Geneva, Switzerland; ⁿCenter for Integrative Genomics, Genopode building, University of Lausanne, 1015 Lausanne, Switzerland; and ^oCentre de Regulació Genòmica, Universitat Pompeu Fabra, E-08003 Barcelona, Spain

Contributed by Janet M. Thornton, January 31, 2007 (sent for review October 20, 2006)

Alternative premessenger RNA splicing enables genes to generate more than one gene product. Splicing events that occur within protein coding regions have the potential to alter the biological function of the expressed protein and even to create new protein functions. Alternative splicing has been suggested as one explanation for the discrepancy between the number of human genes and functional complexity. Here, we carry out a detailed study of the alternatively spliced gene products annotated in the ENCODE pilot project. We find that alternative splicing in human genes is more frequent than has commonly been suggested, and we demonstrate that many of the potential alternative gene products will have markedly different structure and function from their constitutively spliced counterparts. For the vast majority of these alternative isoforms, little evidence exists to suggest they have a role as functional proteins, and it seems unlikely that the spectrum of conventional enzymatic or structural functions can be substantially extended through alternative splicing.

function | human | isoforms | splice | structure

Alternative mRNA splicing, the generation of a diverse range of mature RNAs, has considerable potential to expand the cellular protein repertoire (1–3), and recent studies have estimated that 40–80% of multiexon human genes can produce differently spliced mRNAs (4, 5). The importance of alternative splicing in processes such as development (6) has long been recognized, and proteins coded by alternatively spliced transcripts have been implicated in a number of cellular pathways (7–9). The extent of alternative splicing in eukaryotic genomes has led to suggestions that alternative splicing is key to understanding how human complexity can be encoded by so few genes (10).

The pilot project of the Encyclopedia of DNA Elements (ENCODE) (11), which aims to identify all the functional elements in the human genome, has undertaken a comprehensive analysis of 44 selected regions that make up 1% of the human genome. One valuable element of the project has been the detailing of a reference set of manually annotated splice variants by the GENCODE consortium (12). The annotation by the GENCODE consortium is an extension of the manually curated annotation by the Havana team at The Sanger Institute.

Although a full understanding of the functional implications of alternative splicing is still a long way off, the GENCODE set

has provided us with the material to make an in-depth assessment of a systematically collected reference set of splice variants.

Results

Alternative Splicing Frequency. The GENCODE set is made up of 2,608 annotated transcripts for 487 distinct loci. A total of 1,097 transcripts from 434 loci are predicted to be protein coding. There are on average 2.53 protein coding variants per locus; 182 loci have only one variant, whereas one locus, RP1–309K20.2 (*CPNE1*) has 17 coding variants.

A total of 57.8% of the loci are annotated with alternatively spliced transcripts, although there are differences between target regions chosen manually and those chosen according to the stratified random-sampling strategy (11). The differences stem from gene clusters in the manually selected regions, such as the cluster of 31 loci that code for olfactory receptors in manual pick 9 from chromosome 11 (13). These olfactory receptors are recent in evolutionary origin, have a single large coding exon, and code for a single isoform. This means that although the 0.5% of the human genome that was selected for biological interest has 276 loci, just 52.1% of the loci have multiple variants. In contrast, the regions that were selected in the stratified random-sampling process have fewer loci (158), but 68.7% of the loci have multiple variants (see Fig. 1*a*). This number is toward the higher end of previous estimates but in line with the most recent reports (14).

Analysis of the data suggests that the GENCODE-validated transcripts are an underestimate of the real numbers of alter-

Author contributions: M.L.T., P.L.M., G.A.R., C.Y., M.A., E.B., S.B., R.C., R.G., J.H., H. Hermjakob, D.T.J., T.L., C.A.O., L.P., J.M.T., A.T., and A.V. designed research; M.L.T., P.L.M., A.F., G.A.R., J.J.W., C.Y., P.I.Ó., H. Hegyi, A.G., D.R., J.L., R.L., G.L., M.I.S., J.D.W., P.F., I.R., A.N., W.K., Z.S., M.O., Y.A., H.B., C.H., F.R., A.S., F.D., P.J., S.K., and S.O. performed research; M.L.T., P.L.M., A.F., G.A.R., J.J.W., C.Y., P.I.Ó., H. Hegyi, M.A., A.G., D.R., J.L., R.L., M.I.S., S.E.A., J.D.W., and A.R. analyzed data; and M.L.T., P.L.M., C.Y., P.I.Ó., and A.V. wrote the paper.

The authors declare no conflict of interest.

Abbreviation: TMH, transmembrane helices.

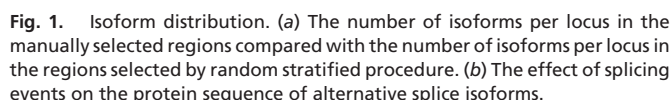
^bTo whom correspondence may be addressed at: Spanish National Cancer Research Centres, c. Melchor Fernández Almagro, 3, E-28029 Madrid, Spain. E-mail: mtress@cniio.es.

^qTo whom correspondence may be addressed. E-mail: helen@ebi.ac.uk.

^rTo whom correspondence may be addressed. E-mail: valencia@cniio.es.

This article contains supporting information online at www.pnas.org/cgi/content/full/070800104/DC1.

© 2007 by The National Academy of Sciences of the USA



A large proportion of the splice isoforms in the data set have identical protein sequences. These coding sequence-identical variants are alternatively spliced in the 5' and 3' untranslated regions and form an interesting subgroup that may be under independent transcriptional control (17). One locus, AF121781.16 (*C21Orf13*), has 11 alternative isoforms, all of which are protein sequence-identical. This is not an isolated case: 230 of the 1,097 isoforms are identical, 25 loci have four or more identical isoforms, and 15% of loci with multiple variants code for nothing but protein sequence-identical isoforms.

Although instances of splicing via functionally interesting mutually exclusive exons (2) were very rare in this set, a number of alternative isoforms are generated from translations of different reading frames. For example, alternative splicing between exons 4 and 5 in isoforms 002 and 003 of locus RP1-309122.1 (*TIMP3*) leads to a frame shift that causes the fifth exon (corresponding the C terminus of the protein) to be read from a different reading frame. There are only three functionally studied examples of overlapping reading frames in humans. One is *INK4a/ARF21* (19), where different transcripts have a coding

Few of the variants coded from overlapping reading frames appeared to be functional. We were able to compare nine transcripts where the human and mouse homologue had conserved exonic structure. If the two alternative reading frames evolve under functional constraints, the mutation rate for all three codon positions should be the same, and both frames should have an identical nonsynonymous substitution rate (K_a , 21). Only one of the nine pairs of transcripts had identical K_a values and equal rates of mutation for each coding position.

Splicing events also lead to alternative isoforms in which it is difficult to predict the resulting membrane topology. In locus AC129929.4 (*TSPAN32*) the principal isoform has four TMH, but the gene also codes for four different splice isoforms that each lose membrane-spanning helices. In isoform 003, the N-terminal helix that acts as both a signal sequence and a membrane anchor (23) is likely to be lost through N-terminal substitution, whereas isoforms 005 and 012 lose the C-terminal TMH. Isoform 014 apparently lacks not just the N-terminal helix, but also the third TMH. This would leave the isoform with two membrane-spanning regions and with the one of the helices oriented in the opposite direction. All these cases must force either a change of structure or polarity.

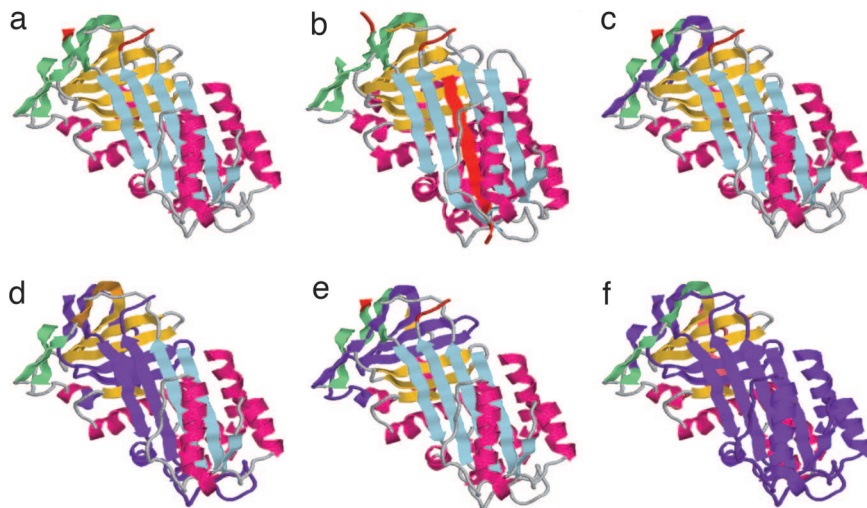


Fig. 3. B serpins. Serpins are protease inhibitors that inactivate their targets after undergoing an irreversible conformational change. (a and b) Serpins exist in an inactivated form (a) that is regarded as being "stressed." Cleavage of the 20-residue RSL region, missing in the structure but with the terminal ends shown in red, causes the RSL region to flip over and fit itself into one of the β -sheets (b, inserted RSL strand in red). This exposes the inhibitory region that inactivates the protease. (c–f) Four splice isoforms from different serpin loci mapped onto the structure of serpinB2 (1×7). The sections deleted/substituted in the isoforms are shown in purple. In each case, it appears that splicing is likely to cause the structure to fold in a substantially different fashion. Given that the complex structure of the inhibitor is vital to its unique function, it is not clear why so many apparently deleterious isoforms would be necessary.

Another locus involved in inflammatory response pathways is IL-4, locus AC004039.4, a cytokine thought to play a role in the development of T helper 2 cells. The GENCODE set contains a single alternative splice variant (isoform 002, IL-4d2) that has a deletion of the second exon, a total of 16 residues. Although the presence of isoform IL-4d2 has yet to be demonstrated in the cell, *in vitro* studies have shown that IL-4d2 retains the ability to bind to IL-4 receptors and acts as a competitive antagonist of IL-4 in monocytes and B cells (33).

The structure of human IL-4 has been well characterized. It is a four-helix bundle with long connecting loops between helices 1 and 2 and 3 and 4 and is held together by three cysteine bridges. The residues coded for by the missing exon coincide with the first of the long loops (see Fig. 4), and to close this loop, a certain amount of structural reorganization relative to the structure of the principal isoform would be necessary. Isoform IL-4d2 has been a favorite target for the homology modeling of splice isoforms, but the size of the gap left by the missing residues and relative inflexibility imposed by the cysteine bridges have hampered predictions. Predicted models have substantially different arrangements of the helices (34–36), and one is even predicted as a knotted structure. Recent results have shown that it is extremely difficult to model even small deletions and insertions with current techniques (37, 38).

Splice Isoforms and Disease. *TAZ* is not the only locus in the set where doubt has been cast on the biological importance of the principal isoform. At least two other loci (AF030876.1, *MECP2* and RP11-247A12.4, *PPP2R4*) are likely to be annotated with the incorrect principal isoform. It has recently been suggested that, because cDNAs for many genes were cloned from tumor samples, the prevalent isoform may well have been coded from a tumor-specific splice variant rather than the mRNA sequence found in normal tissue (39).

For many alternative variants in this set, the mRNA supporting evidence was found exclusively in cancer cell lines, which suggests that the expression of some of these variants may be associated with disease states. It should be borne in mind, however, that tumor lines are overrepresented in cDNA libraries: 26% of the cDNA libraries annotated in the eVOC pathology annotation (40) are annotated as “normal,” whereas 49% are annotated as “tumor.”

There has been abundant recent work associating alternative splicing with stresses incurred by cancer and other disorders (41–43), although in many cases the increase in expression of the aberrant variant may be a side effect of the general breakdown of cellular function rather than part of the instigation process. Indeed the importance of alternative splicing in cancer is such that diagnosis can now be carried out by using isoform-sensitive microarrays based on splice isoform profiles (44).

At least two sets of alternative isoforms in this set are implicated in disease. Isoform 011 from locus AC051649.4 (*TNNT3*) seems to play a role in facioscapulohumeral muscular dystrophy (45), and isoform 006 of locus U52111.6 (*LICAM*) is involved in CRASH syndrome (46).

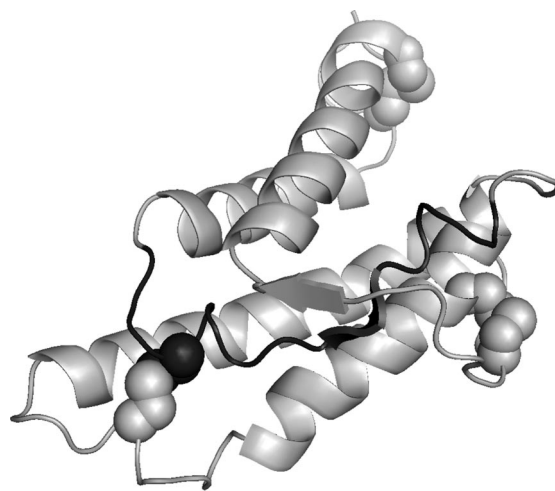


Fig. 4. The difficulty of modeling the structure of isoform IL-4d2. Splice isoform IL-4d2 (isoform 002 from locus AC004039.4) mapped onto structural template 1tl1. The section coded for by the missing second exon is colored in dark gray. The cysteine bridges that stabilize the structure are shown in stick format. The missing exon would leave the flanking residues 30 Å apart, and the cysteine bridges mean that the structure has little room for reorganization. Solutions to the modeling problem would have to break the hydrophobic core of the four-helix bundle or break the cysteine bridges by realigning the helices.

30. Vaz FM, Houtkooper RH, Valianpour F, Barth PG, Wanders RJA (2003) *J Biol Chem* 278:43089–43094.
31. Rao N, Nguyen S, Ngo K, Fung-Leung W-P, (2005) *Mol Cell Biol* 25:6521–6532.
32. Jensen LE, Whitehead AS (2001) *J Biol Chem* 276:29037–29044.
33. Arinobu Y, Atamas SP, Otsuka T, Niiro H, Yamaoka K, Mitsuyasu H, Niho Y, Hamasaki N, White B, Izuhara K (1999) *Cell Immunol* 191:161–167.
34. Zav'yalov VP, Denesyuk AI, White B, Yurovsky VV, Atamas SP, Korpela T (1997) *Immunol Lett* 58:149–152.
35. Furnham N, Ruffle S, Southan C (2003) *Proteins* 54:596–608.
36. Wen F, Li F, Xia H, Lu X, Zhang X, Li Y (2004) *Trends Genet* 20:232–236.
37. Ginalski K (2006) *Curr Op Struct Biol* 16:172–177.
38. Tress ML, Ezkurdia I, Graña O, López G, Valencia A (2005) *Proteins* 61:27–45.
39. Roy M, Xu Q, Lee C (2005) *Nucleic Acids Res* 33:5026–5033.
40. Law DJ, Labut EM, Adams RD, Merchant JL (2006) *Nucleic Acids Res* 34:1342–1350.
41. Kishore S, Stamm S (2006) *Science* 311:230–232.
42. Ottenheijm CAC, Heunks LMA, Hafmans T, van der Ven PFM, Benoist C, Zhou H, Labeit S, Granzier HL, Dekhuijzen PNR (2006) *Am J Respir Crit Care Med* 173:527–534.
43. Brinkman BM (2004) *Clin Biochem* 37:584–594.
44. Kelso J, Visagie J, Theiler G, Christoffels C, Bardien-Kruger S, Smedley D, Otgaar D, Greyling G, Jongeneel V, McCarthy M, et al. (2003) *Genome Res* 13:1222–1230.
45. Jacob J, Haspel J, Kane-Goldsmith N, Grumet MJ (2002) *Neurobiol* 51:177–189.
46. Gabellini D, D'Antona G, Moggio M, Prella A, Zecca C, Adami R, Angeletti B, Ciscato P, Pellegrini MA, Bottinelli R, et al. (2006) *Nature* 439:973–977.
47. Xing Y, Lee C (2006) *Nat Rev Genet* 7:499–509.
48. Hoffmann R, Valencia A (2005) *Bioinformatics* 21:252–258.
49. Bendtsen JD, Nielsen H, von Heijne G, Brunak S (2004) *J Mol Biol* 16:783–795.
50. Bernsel A, von Heijne G (2005) *Prot Sci* 14:1723–1728.
51. Martelli PL, Fariselli P, Casadio R (2003) *Bioinformatics* 19:1205–1211.
52. Viklund H, Elofsson A (2004) *Prot Sci* 13:1908–1917.
53. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) *Nucleic Acids Res* 25:3389–3402.
54. Slater GS, Birney E (2005) *BMC Bioinformatics* 6:31.